Statistical anaphora resolution in biomedical texts

Caroline Gasperin **Ted Briscoe**
Computer Laboratory
University of Cambridge
Cambridge, UK
{cvg20,ejb}@cl.cam.ac.uk

Abstract

This paper presents a probabilistic model for resolution of non-pronominal anaphora in biomedical texts. The model seeks to find the antecedents of anaphoric expressions, both coreferent and associative ones, and also to identify discourse-new expressions. We consider only the noun phrases referring to biomedical entities. The model reaches state-of-the art performance: 56-69% precision and 54-67% recall on coreferent cases, and reasonable performance on different classes of associative cases.

1 Introduction

Inspired by Ge et al. (1998) probabilistic model for pronoun resolution, we have developed a model for resolution of non-pronominal anaphora in biomedical texts.

The probabilistic model results from a simple decomposition process applied to a conditional probability equation that involves several parameters (features). The decomposition makes use of Bayes' theorem and independence assumptions, and aims to decrease the impact of data sparseness on the model, so that even small training corpora can be viable. The decomposed model can be understood as a more sophisticated version of the naive-Bayes algorithm, since we consider the dependence among some of the features instead of full independence as in naive Bayes. Probabilistic models can return a confidence measure (probability) for each decision they make, while decision trees, for example, cannot. Another advantage of this type of model is the fact that they consider the prior probability of each class, while other machine-learning techniques such as SVMs and neural networks do not.

Our model seeks to classify the relation between an anaphoric expression and an antecedent candidate as coreferent, associative or neither. It computes the probability of each pair of anaphor and candidate for each class. The candidate with the highest overall probability for each class is selected as the antecedent for that class, or no antecedent is selected if the probability of no relation overcomes the positive probabilities; in this case, the expression is considered to be new to the discourse.

Coreferent cases are the ones in which the anaphoric expression and its antecedent refer to the same entity in the real world, as below:

(1) "The expression of **reaper** has been shown ... **the gene** encodes ..."

Associative cases are the ones in which the anaphoric expression and its antecedent refer to different entities, but where the antecedent plays a role in defining the meaning of the anaphoric expression, as in Example 2:

(2) "**Drosophila gene Bok** interacts with ... expression of **Bok protein** promotes apoptosis ..."

Discourse new cases usually consist of the first mention of an entity in the text, so no antecedent can be found for it.

We have focused on the biomedical domain for two reasons. Firstly, there is a vast demand from the biomedical field for information extraction efforts (which require NLP processing, including

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 257–264
Manchester, August 2008

anaphora resolution), in order to process the exponentially increasing number of journal publications, which are the major source of novel knowledge to be extracted and condensed into domain databases for easy access. Secondly, anaphora resolution can benefit from the several sources of refined semantic knowledge that are not commonly available for other domains, such as biomedical databases, ontologies, and terminologies.

In the next section, we describe the anaphoric relations that we found in biomedical texts, which we are considering for the resolution process. In Section 3 we describe our probabilistic model, and in Section 4 the corpus created for training it. In Section 5 we present and discuss the results achieved by the model, and compare it with a baseline system and a decision-tree-based system.

2 Anaphora cases in biomedical text

Biomedical texts differ from other genres of text (e.g. newswire, fiction) in several points. Different types of NPs have a particular distribution in biomedical articles. For example, pronouns are very rare, accounting for a very small percentage of the noun phrases[1]; on the other hand, proper names occur very often, given the frequent mention of gene and protein names and the names of other biomedical entities. A system for anaphora resolution in the biomedical domain can benefit from focusing on the most common types of noun phrases, that is, non-pronominal ones.

In biomedical articles, the reader needs background knowledge to understand the underlying relation between the entities mentioned in the text in order to understand the text. For instance, in Example 2 the reader is expected to know that a gene encodes a protein (which usually carries its name), so that he/she can capture the anaphoric relation and understand the sentence. This aspect emphasises the need for semantic information as part of the anaphora resolution process.

Another aspect affecting the anaphoric relations in biomedical texts are the writing conventions adopted in the biomedical domain to distinguish between the mention of a gene name and the mention of the protein encoded by that gene. The most usual convention is writing gene names with lowercase italicised letters and protein names with non-italicised uppercase letters. The existence of such conventions allows for constructions where the reader keeps the conventions in mind to understand the text, as below.

(3) "Drosophila has recently been shown also to have a CED-4/Apaf-1 homolog, named **Dark/HAC-1/Dapaf-1**. ... Loss of function mutations in ***dark/hac-1/dapaf-1*** result in ..."

Among the associative cases present in biomedical text, we were able to distinguish two main types of relations. The first, which we call "biotype" relations, are associative relations between biomedical entities with different semantic types, which we call biotypes (e.g. gene, gene product, part of gene). This is the case of Example 2 and 3[2]. If we take into account the specific biotype of the entities that are involved in a biotype relation, it is possible to determine a WordNet-like semantic relation behind the anaphora relation. For example, a biotype relation between a 'gene' and a 'variant' of gene can be considered an hyponymy relation, the relation between a 'gene' and a transcript (part of gene) can be seen as a meronymy relation.

The second type of associative relation is more common to other domains as well, we call it "set-member" relation. It consists of cases where the anaphor refers to a set that contains the antecedent, or vice-versa, as in Examples 4 and 5.

(4) ... **ced-4** and **ced-9** ... **the genes** ...

(5) ... **the mammalian anti-apoptotic protein Bcl-2** ... **Bcl-2 family** ...

3 The resolution model

Our probabilistic model aims to find the closest coreferent and/or associative antecedent for all non-pronominal NPs that refer to biomedical entities. Among non-pronominal NPs we distinguish proper names, definite NPs (e.g. "the gene"), demonstrative NPs (e.g. "this gene"), indefinite NPs (e.g. "a gene"), quantified NPs (e.g. "four genes", "some genes") and other NPs.

We consider the three classes of anaphoric relations mentioned above: coreferent, associative biotype, and associative set-member.

We have chosen 11 features to describe the anaphoric relations between two noun phrases. The features are presented in Table 1. Most features are domain-independent, while one, gp, is

[1] About 3% of the noun phrases according to the corpus presented in Section 4.

[2] Associative relations between proper names are not known to happen in other domains, and are made possible in the biomedical domain given the existence of naming conventions.

specific for the biomedical domain. Our feature set covers the basic aspects that influence anaphoric relations: the form of the anaphor's NP, string matching, semantic class matching, number agreement, and distance.

Given these features, for each antecedent candidate a of an anaphor A, we compute the probability P of an specific class of anaphoric relation C between a and A. P is defined as follows:

$$P(C = \text{'class'}|f_A, f_a, hm_{a,A}, hmm_{a,A}, mm_{a,A}, num_{a,A}, sr_a, bm_{a,A}, gp_{a,A}, d_{a,A}, dm_{a,A})$$

For each pair of a given anaphor and an antecedent candidate we compute P for C='coreferent', C='biotype', and C='set-member'. We also compute C='none', that represents the probability of no relation between the NPs.

We decompose the probability P and assume independence among some of the features in order to handle the sparseness of the training data. In the following equations, we omit the subscripted indexes of the relational features for clarity.

$$P(C|f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm) = \frac{P(C)P(f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm|C)}{P(f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm)} \quad (1)$$

Equation 1 is obtained by applying Bayes' theorem to the initial equation. $P(C)$ is the prior probability of each class, it will encode the distribution of the classes in the training data. As the denominator contains feature values that change according to the candidate being considered, we cannot eliminate it in the usual fashion, so we keep it in order to normalise P across all candidates. From this equation, we then selectively apply the chain rule to both numerator and denominator until we get to the following equation:

$$= \frac{\begin{array}{c} P(C)\ P(f_A|C)\ P(f_a|C, f_A)\ P(d, dm|C, f_A, f_a) \\ P(sr|C, f_A, f_a, d, dm)\ P(bm, gp|C, f_A, f_a, d, dm, sr) \\ P(num|C, f_A, f_a, d, dm, sr, bm, gp) \\ P(hm, hmm, mm|C, f_A, f_a, d, dm, sr, bm, gp, num) \end{array}}{\begin{array}{c} P(f_A)\ P(f_a|f_A)\ P(d, dm|f_A, f_a) \\ P(sr|f_A, f_a, d, dm)\ P(bm, gp|f_A, f_a, d, dm, sr) \\ P(num|f_A, f_a, d, dm, sr, bm, gp) \\ P(hm, hmm, mm|f_A, f_a, d, dm, sr, bm, gp, num) \end{array}} \quad (2)$$

Following the decomposition, we eliminate some of the dependencies among the features that we consider unnecessary[3]. We consider that the lexical features hm, hmm, and mm are not dependent on distance d or dm, nor on sr, gp or num, so:

$$P(hm, hmm, mm|C, f_A, f_a, d, dm, sr, bm, gp, num) \propto P(hm, hmm, mm|C, f_A, f_a, bm)$$

We model num as independent from d, dm, sr, bm, and gp, so:

$$P(num|C, f_A, f_a, d, dm, sr, bm, gp) \propto P(num|C, f_A, f_a)$$

We also assume the semantic features bm, and gp as independent from all features but C:

$$P(bm, gp|C, f_A, f_a, d, dm, sr) \propto P(bm, gp|C)$$

We also assume sr to be independent of f_A and f_a:

$$P(sr|C, f_A, f_a, d, dm) \propto P(sr|C, d, dm)$$

The final equation then becomes:

$$P(C|f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm) = \frac{\begin{array}{c} P(C)\ P(f_A|C)\ P(f_a|C, f_A)\ P(d, dm|C, f_A, f_a) \\ P(sr|C, d, dm)\ P(bm, gp|C)\ P(num|C, f_A, f_a) \\ P(hm, hmm, mm|C, f_A, f_a, bm) \end{array}}{\begin{array}{c} P(f_A)\ P(f_a|f_A)\ P(d, dm|f_A, f_a) \\ P(sr|d, dm)\ P(bm, gp)\ P(num|f_A, f_a) \\ P(hm, hmm, mm|f_A, f_a, bm) \end{array}} \quad (3)$$

4 Training

There are very few biomedical corpora annotated with anaphora information, and all of them are built from paper abstracts (Cohen et al., 2005), instead of full papers. As anaphora is a phenomenon that develops through a text, we believe that short abstracts are not the best source to work with and decided to concentrate on full papers.

In order to collect the statistics to train our model, we have manually annotated anaphoric relations between biomedical entities in 5 full-text articles (approx. 33,300 words)[4], which are part of the Drosophila molecular biology literature. The corpus and annotation process are described in (Gasperin et al., 2007). To the best of our knowledge, this corpus is the first corpus of biomedical full-text articles to be annotated with anaphora information.

[3] For brevity, we only show this process for the numerator, although the same is assumed for the denominator.

[4] Corpus available via the FlySlip project website http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip

Feature	Possible values
f_A	Form of noun phrase of the anaphor A: 'pn', 'defnp', 'demnp', 'indefnp', 'quantnp', or 'np'.
f_a	Form of noun phrase of the antecedent candidate a: same values as for f_A.
$hm_{a,A}$	Head-noun matching: 'yes' if the anaphor's and the candidate's head nouns match, 'no' otherwise.
$hmm_{a,A}$	Head-modifier matching: 'yes' if the anaphor's head noun matches any of the candidate's pre-modifiers, or vice-versa, 'no' otherwise.
$mm_{a,A}$	Modifier matching: 'yes' if anaphor and candidate have at least one head modifier in common, 'no' otherwise.
$num_{a,A}$	Number agreement: 'yes' if anaphor and candidate agree in number, 'no' otherwise.
$sr_{a,A}$	Syntactic relation between anaphor and candidate: 'none', 'apposition', 'subj-obj', 'pp', and few others.
$bm_{a,A}$	Biotype matching: 'yes' if anaphor's and candidate's biotype (semantic class) match, 'no' otherwise.
$gp_{a,A}$	is biotype *gene* or *product*? 'yes' if the anaphor biotype or candidate biotype is *gene* or *product*, 'no' otherwise. This feature is mainly to distinguish which pairs can hold biotype relations.
$d_{a,A}$	Distance in sentences between the anaphor and the candidate.
$dm_{a,A}$	Distance in number of entities (markables) between the anaphor and the candidate.

Table 1: Feature set

Before annotating anaphora, we have preprocessed the articles in order to (1) tag gene names, (2) identify all NPs, and (3) classify the NPs according to their domain type, which we call biotype. To tag all gene names in the corpus, we have applied the gene name recogniser developed by Vlachos et al. (2006). To identify all NPs, their subconstituents (head, modifiers, determiner) and broader pre- and post-modification patterns, we have used the RASP parser (Briscoe et al., 2006). To classify the NPs according to their type in biomedical terms, we have adopted the Sequence Ontology (SO)[5] (Eilbeck and Lewis, 2004). SO is a fine-grained ontology, which contains the names of practically all entities that participate in genomic sequences, besides the relations among these entities (e.g. is-a, part-of, derived-from relations). We derived from SO seven biotypes to be used to classify the entities in the text, namely: "gene", "gene product", "part of gene", "part of product", "gene variant", "gene subtype", and "gene supertype". We also created the biotype "other-bio" to be associated with noun phrases that contain a gene name (identified by the gene name recogniser) but whose head noun does not fit any of the other biotypes. All NPs were tagged with their biotypes, and NPs for which no biotypes were found were excluded.

The gene-name tags, NP boundaries and biotypes resulting from the preprocessing phase were revised and corrected by hand before the anaphoric relations were annotated.

For each biotyped NP we annotated its closest coreferent antecedent (if found) and its closest associative antecedent (if found), from one of the associative classes. From our annotation, we can infer coreference chains by merging the coreferent links between mentions of a same entity.

The annotated relations, and the features derived from them, are used as training data for the probabilistic model above. We have also considered negative training samples, which result from the absence of an anaphoric relation between a NP that precedes an anaphoric expression and was not marked as its antecedent (nor marked as part of the same coreference chain of its antecedent). The negative samples outnumber considerably the number of positive samples (annotated cases). Table 2 presents the distribution of the cases among the classes of anaphora relations.

We note that around 80% of the definite NPs are anaphoric in our corpus, instead of the 50% presented in (Vieira and Poesio, 2000) for newspaper texts. Nearly all demonstrative NPs (93%) are anaphoric. More than 70% of the proper names take part in coreference relations, as they inherently refer to a specific named entity, but nevertheless 5% of them take part in associative biotype relations, due to the fact that a gene and the protein it synthesizes usually share the same name. 44% of quantified NPs take part in set-member relations, as they usually refer to more than one entity. Finally 51% of indefinite NPs are discourse new.

To balance the ratio between positive and negative training samples, we have clustered the negative samples and kept only a portion of each cluster, proportional to its size. All negative samples that have the same values for all features are grouped together (consequently, a cluster is formed by a set of identical samples) and only $\frac{1}{10}$ of each cluster members is kept, resulting in 85,314 negative samples. This way, small clusters (with less than 10 members), which

[5] http://www.sequenceontology.org/

Class/NPs	pn	defnp	demnp	indefnp	quantnp	other np	Total
coreferent	689	429	70	40	54	396	1678
biotype	43	102	3	8	4	114	274
set-member	151	126	26	14	68	158	543
discourse new	63	107	0	72	38	156	436
none	873,731						

Table 2: Training instances, according to anaphoric class and to NP form

are likely to represent noisy samples (similar to positive ones), are eliminated, and bigger clusters are shrunk; however the shape of the distribution of the negative samples is preserved. For example, our biggest cluster (feature values are: f_A='pn', f_a='pn', hm='no', hmm='no', mm='no', bm='yes', gp='yes', num='yes', sr='none', d='16<', dm='50<') with 33,998 instances is reduced to 3,399 – still considerably more numerous than any positive sample.

Other works have used a different strategy to reduce the imbalance between positive and negative samples (Soon et al., 2001; Ng and Cardie, 2002; Strube et al., 2002), where only samples composed by a negative antecedent that is closer than the annotated one are considered. We compare the performance of both strategies in Section 5.1 and show that ours is more effective. The higher the number of negative samples, the higher the precision of the resolution, but the lower the recall.

5 Results

Given the small size of our corpus, we did not hold out a test set. Instead, we have measured the average performance achieved by the model in a 10-fold cross-validation setting, using the whole of the annotated corpus.

We consider as antecedent candidates all noun phrases that precede the anaphor. For a given anaphor, we first select as antecedent according to each anaphora class the candidate with the highest value for P for that class. We also compute $P(C$='none') for all candidates. If $P(C$='coreferent') > $P(C$='none') for the selected coreferent antecedent, it is kept as the resulting antecedent. The same is tested for the selected associative antecedent with the highest probability, independent of the type of associative class. For set-member cases, where an anaphor can have multiple antecedents, if more than one candidate has an equally high probability, all these candidates are kept. When no coreferent or associative antecedent is found (or when $P(C$='none') is higher on both cases) the anaphor is classified as discourse new.

Table 3 presents the performance scores we achieved for each anaphora class. The first column, 'perfect', shows the result of a strict evaluation, where we consider as correct all pairs that match exactly an antecedent-anaphor pair in the annotated data. On the other hand, column 'relaxed' treats as correct also the pairs where the assigned antecedent is not the exact match in the annotated data but is coreferent with it.

It is clear that the results for coreferent cases are much better than for associative cases, but the latter are known to be more challenging. On top of that, the 'relaxed' column shows considerable improvements in comparison to 'perfect'. That means that several anaphors are being linked to the correct coreference chain, despite not being linked to the closest antecedent. This happens mainly in cases where there is no string matching between the closest antecedent and the anaphor, causing an earlier mention of the same entity with matching head and/or modifiers to get higher probability. We believe we can approximate 'perfect' to 'relaxed' results if we extend the string matching features to represent the whole coreference chain, that is, consider a positive matching when the anaphor match any of the elements in a chain, similarly to the idea presented in (Yang et al., 2004).

We believe that the lower overall performance for associative cases is due to the difficulty of selecting features that capture all aspects involved in associative relations. Our set of features is clearly failing to cover some of these aspects, and a deeper feature study should be the best way to boost the scores. However, despite lower, these performance

Class	Perfect			Relaxed		
	P	R	F	P	R	F
coreferent	56.3	54.7	55.5	69.4	67.4	68.3
biotype	28.5	35.0	31.4	31.2	37.9	34.2
set-member	35.4	38.2	36.7	38.5	41.5	40.0
discourse new	44.3	53.4	48.4	44.3	53.4	48.4

Table 3: Performance of the probabilistic model

scores are higher that the ones from previous approaches for newspaper texts, which used for instance the WordNet (Poesio et al., 1997) or the Internet (Bunescu, 2003) as source of semantic knowledge.

We have analysed our features and observed that the string matching features hm, hmm, and mm, the number agreement feature num, biotype matching bm, and distance in markables dm are the core features and achieve reasonable performance. However, f_A and f_a play an important role, they increase the precision of coreferent cases and boost considerably the performance of the associative ones. This is due to the different distribution of NP types across the relations as shown is Table 2. The remaining features focused on specific cases: gp improved biotype recall, by boosting the probability of a biotype relation when anaphor or candidate had specific biotypes; and sr improved precision and recall of coreferent cases.

Table 4 shows the 'perfect' performance scores according to each class of NP. The resolution of proper names achieves the highest scores among all types of NPs for most classes. That is due to their limited structure, since proper names usually do not have elaborated pre-modification or modification at all, so our string matching features carried simpler patterns for these NPs. Indefinite and quantified NPs achieved the lowest scores for coreferent cases, since the highest percentage of training instances for these NPs are not coreferent (as seen in Table 2). Indefinite NPs, as expected, have the best scores for discourse new cases.

5.1 Comparing to other approaches

We have tried training our probabilistic model using a different strategy than the one described in Section 4 for selecting negative samples. This strategy consists of selecting only the negative samples that occur between the anaphor and its coreferent antecedent, not considering candidates that are further away than the antecedent. This strategy was first used for anaphora resolution by Soon et al. (2001). Column 'prob+closest' on Table 5 shows the performance scores. In our dataset, this strategy was able to reduce the number of negative samples to about $\frac{1}{3}$ of its size, while our strategy reduces it to $\frac{1}{10}$. The larger number of negative samples increases the precision scores and reduces the recall scores for all positive classes, while the opposite happens for the negative class, which defines the discourse new scores. We reckon that the considerable drop on recall numbers for the associative cases would make the system less viable, while the low precision for discourse new cases shows that many anaphoric cases are left unresolved. We view our strategy, based on the clustering of negative samples and consecutive cluster size reduction, to be more effective at proportionally eliminating negative samples that are less frequent and that are more likely to be noisy.

We compare our model to a rule-based baseline system that we have previously developed. The baseline system (Gasperin, 2006) for each anaphoric expression: 1) selects as coreferent antecedent the closest preceding NP that has the same head noun, same biotype and agrees in number with the anaphor, and 2) selects as associative antecedent the closest preceding NP that has the same head noun, same biotype but disagrees in number with the anaphor, or that has the same head noun or a modifier matching the anaphor head (or vice-versa) or matching modifiers, agrees in number but has different biotypes. The baseline system does not distinguish between different types of associative cases, although it aims to cover biotype and set-member cases. If no antecedent that matches these criteria is found, the anaphor is considered discourse new. Column 'baseline' on Table 5 shows the performance scores for the baseline system. The scores for coreferent cases are reasonable, despite being below our probabilistic model, while the scores for associative cases, especially recall, are considerably lower. The baseline system relies on some sort of string matching between anaphor and antecedent, and is not able to infer a relation between expressions when the matching does not happen. That is one of the main aspects that the probabilistic system tries to overcome by weighting the contribution of all features.

We also compared our model to a system based on decision trees, since this approach has been taken by several corpus-based anaphora resolution systems (Soon et al., 2001; Ng and Cardie, 2002; Strube et al., 2002). We have induced a decision tree using the C4.5 algorithm implemented in the Weka tool (Witten and Frank, 2005); we have used the same features used for our probabilistic model. We selected as the antecedent the candidate which is the closest one to the anaphor for which a class other then 'none' is assigned by the decision tree. The 'perfect' and 'relaxed' scores for C4.5 are pre-

Class	coreferent			biotype			set-member			discourse new		
	P	R	F	P	R	F	P	R	F	P	R	F
pn	77.5	71.9	74.6	26.8	25.5	26.1	53.7	65.7	59.1	35.1	59.3	44.1
defnp	48.0	47.3	47.6	26.3	28.1	27.2	29.2	26.1	27.6	38.8	51.8	44.4
demnp	57.8	48.5	52.8	-	-	-	71.4	57.6	63.8	-	-	-
indefnp	27.0	34.2	30.2	14.2	12.5	13.3	21.0	28.5	24.2	63.4	54.7	58.8
quantnp	11.2	12.9	12.0	-	-	-	28.5	37.6	32.5	37.1	34.2	35.6
other np	41.3	41.4	41.4	30.9	48.2	37.7	19.3	19.4	19.4	49.7	56.0	52.6

Table 4: Performance of the probabilistic model ('perfect') per NP form

sented in the last two columns of Table 5. We note that the difference between 'perfect' and 'relaxed' scores is not as large as for our probabilistic model; that shows that decision trees are more often getting even the coreference chain wrong, not just the closest antecedent. We assume this is due to the lack of ranking among the candidates, since we adopt the default strategy of selecting the closest candidate that gets a positive class according to the tree.

The main disadvantage of both the baseline and decision tree systems when compared to the probabilistic model, besides the lower performance, is that they do not provide a probability assigned to each decision they make, which makes it impossible to know how confident the model is for different cases and to take advantage of that information to improve the system. This aspect also makes it difficult to develop a consistent strategy for returning multiple antecedents for set-member cases, since there is no obvious way to do it.

6 Related work

We are not aware of any learning-based system which has dealt with coreferent as well as associative cases of anaphora.

Viera and Poesio (2000) have developed a heuristic-based system for coreferent and associative anaphora resolution of definite NPs in newspaper texts, and have reached 62% recall and 83% precision for direct anaphora (coreferent cases with same head noun), but poor performance for bridging cases (associative cases + coreferent cases with different head nouns) using WordNet as souce of semantic knowledge.

Ng and Cardie (2002), extending the work of Soon et al. (2001), have developed a machine-learning system just for coreference resolution of all types of NPs, also on newspaper texts. Their best results were 64.2% recall and 78.0% precision.

The best-known system to resolve anaphora in the biomedical domain is the work of Castaño et al. (2002), who developed a salience-based system for resolution of coreferent cases. It seeks to resolve pronouns and nominal (which they call sortal) anaphora. As a source of semantic knowledge, they have used the UMLS Semantic Network types[6], which they report to be too coarse grained, and assume that a finer-grained typing strategy would help to increase the precision of the resolution system. They achieved 74% precision and 75% recall on a very small test set.

Yang et al. (2004) implemented a machine-learning approach to coreference resolution similar to Ng and Cardie's, and evaluated it on a portion of the GENIA corpus, which is tagged with semantic information based on the GENIA Ontology[7]. They achieved recall of 80.2% and precision of 77.4%.

Both the Castaño et al. and Yang et al. systems have been developed based on abstracts of biomedical articles, instead of full-text articles, which involve only restricted use of anaphora.

7 Conclusion and future work

We have presented a probabilistic model for resolving anaphoric NPs in biomedical texts. We are not aware of previous works which have dealt with coreferent and associative anaphora in the biomedical domain. Our model, despite being simple and being trained on a very small corpus, coped well with its task of finding antecedents for coreferent and associative cases of anaphora, and was able to achieve state-of-the-art performance. It has outperformed our baseline system and a decision-tree-based system using the same set of features.

Our model returns a probability for each classification it makes, and this can be used as a confidence measure that can be exploited to improve the system itself or by external applications.

Due to our small corpus, we had to limit the

[6] http://www.nlm.nih.gov/research/umls/

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

Class	Prob+Closest			Baseline			C4.5			C4.5 relaxed		
	P	R	F	P	R	F	P	R	F	P	R	F
coreferent	66.2	50.0	56.9	47.0	57.6	51.8	49.6	58.1	53.5	52.7	61.6	56.8
biotype	31.1	10.1	15.2	28.6	10.7	15.6	21.7	28.5	24.6	22.9	29.9	26.0
set-member	46.3	17.5	25.4				28.5	31.3	29.8	30.4	33.3	31.8
discourse new	31.3	88.1	46.2	37.3	30.2	33.4	48.5	32.5	38.9	48.5	32.5	38.9

Table 5: Performance of other models

number and the complexity of the features we use, since the more features, the more sparse the data, and the more training data needed. However, we aim to expand the feature set with more fine-grained features.

Our current work involves using the probabilistic model presented here as part of an active learning framework. The confidence of the model for each decision (probability) is used to selectively gather more samples from unlabelled data and iteratively improve the performance of the system.

The probabilistic model is intended to replace the baseline system in a tool designed to help biology researchers to curate scientific papers (Karamanis et al., 2008).

Acknowledgements

This work is part of the BBSRC-funded FlySlip project. Caroline Gasperin is funded by a CAPES award from the Brazilian government.

References

Briscoe, Edward J., John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of ACL-COLING 06*, Sydney, Australia.

Bunescu, Razvan. 2003. Associative anaphora resolution: A web-based approach. In *Proceedings of EACL 2003 - Workshop on The Computational Treatment of Anaphora*, Budapest.

Castaño, José, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of International Symposium on Reference Resolution for NLP 2002*, Alicante, Spain.

Cohen, K. Bretonnel, Lynne Fox, Philip Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processsing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, Detroit.

Eilbeck, Karen and Suzanna E. Lewis. 2004. Sequence ontology annotation guide. *Comparative and Functional Genomics*, 5:642–647.

Gasperin, Caroline, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC 2007*, Lagos, Portugal.

Gasperin, Caroline. 2006. Semi-supervised anaphora resolution in biomedical texts. In *Proceedings of BioNLP'06*, New York.

Ge, Niyu, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora - COLING-ACL'98*, Montreal, Canada.

Karamanis, Nikiforos, Ruth Seal, Ian Lewin, Peter McQuilton, Andreas Vlachos, Caroline Gasperin, Rachel Drysdale, and Ted Briscoe. 2008. Natural language processing in aid of flybase curators. *BMC Bioinformatics*, 9(193).

Ng, Vincent and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, Philadelphia.

Poesio, Massimo, Renata Vieira, and Simone Teufel. 1997. Resolving bridging descriptions in unrestricted texts. In *Proceedings of the Workshop on Operational Factors In Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Strube, Michael, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the EMNLP 2002*, Philadelphia.

Vieira, Renata and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579.

Vlachos, Andreas and Caroline Gasperin. 2006. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of BioNLP at HLT-NAACL 2006*, pages 138–145, New York.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.

Yang, X., J. Su, G. Zhou, and C. L. Tan. 2004. An NP-cluster based approach to coreference resolution. In *Proceedings of COLING 2004*, Geneva, Switzerland.